# LEARNING TO PREDICT WHERE THE CHILDREN WITH ASD LOOK

*Huiyu Duan, Guangtao Zhai, Xiongkuo Min, Yi Fang, Zhaohui Che, Xiaokang Yang, Cheng Zhi, Hua Yang and Ning Liu*

Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai, China
Email: {huiyuduan, zhaiguangtao, minxiongkuo, yifang}@sjtu.edu.cn

## ABSTRACT

As is known to us, people with Autism Spectrum Disorder (ASD) have atypical visual attention towards stimuli. Learning the visual attention of people especially, children, with ASD contribute to related research in the field of medicine and psychology. In this paper, we first construct a saliency prediction for children with autism (SPCA) database, which is the first of its kind and consists of 500 images and the corresponding eye tracking data collected from 13 different children with ASD. We compare the performance of five state-of-the-art deep neural networks (DNN)-based saliency prediction approaches with their original networks and the fine-tuned networks on our database. We predict the atypical visual attention of children with ASD for the first time and get the best saliency prediction results for individuals with ASD so far.

***Index Terms*—** Children with ASD, visual attention, saliency model, DNN, database

## 1. INTRODUCTION

People with Autism Spectrum Disorder (ASD) show disparate attention in real life, especially in social communication activities. Learning such atypical visual attention could help us understand ASD better. Previous studies have obtained many conclusions in this research field. The study of Dawson *et al.* [1] confirmed that individuals with ASD show reduced attention to faces or other social stimuli but pay more attention to objects. Osterling and Dawson [2] have shown that individuals with autism have an reduced social and joint attention behaviors.

Eye movements encode rich information about attention, oculomotor control and psychological factors of an individual. Thus from the eye movements of people with autism, we can characterize ASD traits. This could even help with ASD's diagnosis. Tseng *et al.* [3] analyzed gaze pattern of watching short video clips. They combined gaze pattern and low-level features together and showed the advantages of identifying specific disorders with the incorporation of attention-related features. But they did not consider the high-level semantic and social information. With the help of machine learning, high-level semantic features can be extracted easily, so the difference between people with autism and healthy individuals could be investigated more accurately than before. Liu *et al.* [4] proposed a machine learning method to classify children with ASD and control groups based on the gaze patterns of children with ASD in a face recognition task. Wang *et al.* [5] quantified atypical visual attention in ASD across multiple levels of features of images for the first time. The main limitations of the methods in [4, 5] are the handcrafted features and requirement of manual labeled objects of interest.

With the advent of deep neural networks (DNN), research related to visual attention of individuals with autism has made great progress recently. Jiang *et al.* [6] fine-tuned one saliency prediction algorithm using the fixation data of people with autism and obtained better performance in the classification of individuals with autism and healthy controls. They also tried to predict the difference map between two groups when they look at an image. In this paper, we go one step further, and design specific saliency model for ASD. On one hand, with the help of good saliency models designed for ASD, we can diagnose ASD better. On the other hand, the specific saliency models can also help us design the specialized contents such as textbook, so that the people with ASD can grasp the contents more easily.

Owing to the progress of DNN and the large annotated datasets, saliency prediction has achieved great improvements [7, 8, 9, 10] and performed better than traditional methods [11, 12, 13]. With sufficient training data, DNN-based models have overwhelming advantages of self-adjustment than traditional models in specific application fields. When the training samples are inadequate, we can also use the fine-tuning methods. Therefore, in this paper, we transfer five state-of-the-art DNN-based saliency prediction models to predict where the children with ASD look and get the saliency models for autistic. To accomplish this goal, we establish a saliency prediction for children with autism (SPCA) database including 500 images. Our database will be released to facilitate further research and we will continue to expand the scale of the database such that it could be used to train the network. But we can only use this database to fine-tune the network so far. We collected eye tracking data of children with autism rather than adults because it is more important for early diagnosis and intervention treatment of children with autism.

The remainder of this paper is arranged as follows. In Sec-

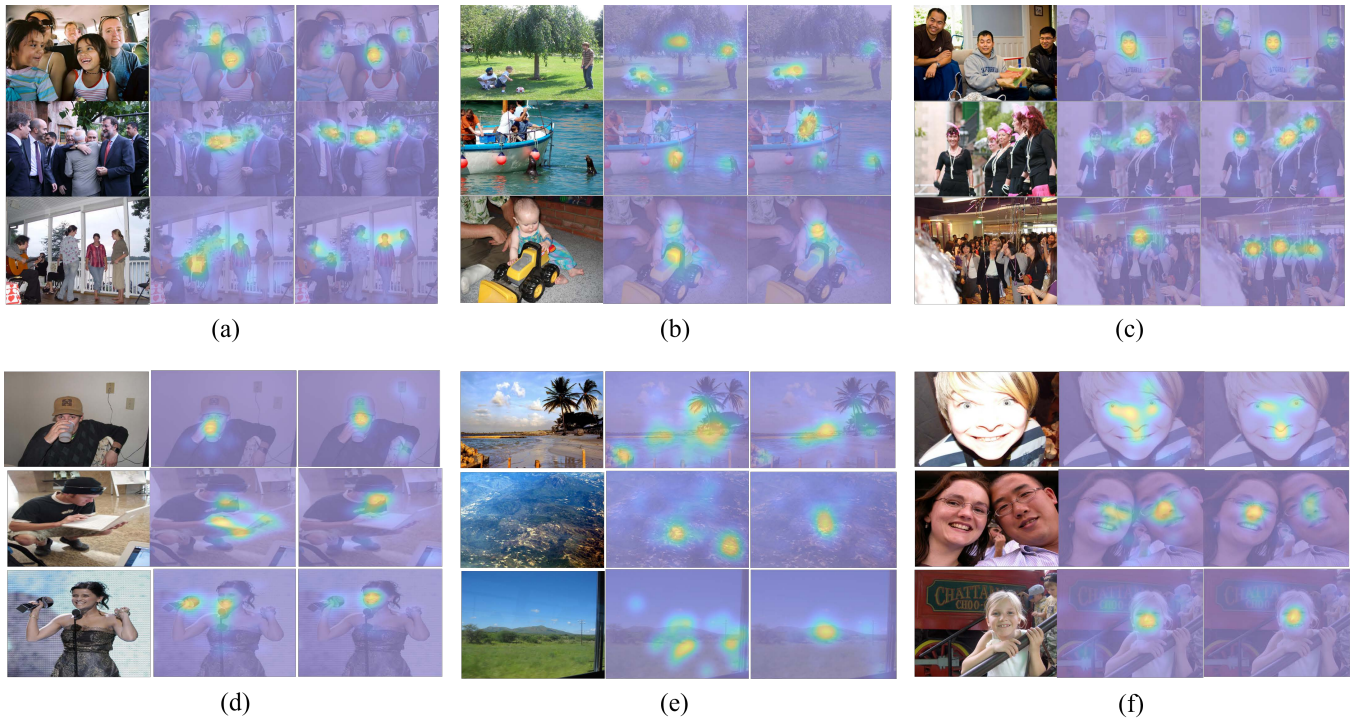

**Fig. 1**. Comparison between autistic's visual attention map and healthy controls' visual attention map (three columns of each subfigure from left to right are sample image, heat map of autistic and heat map of healthy controls, respectively.) (a) joint attention difference. (b) objects or animals bias. (c) center bias. (d) hand bias. (e) against center bias of natural scenes. (f) human faces with similar visual attention map.

tion 2, we introduce the eye-tracking experiments and analyze the collected eye-tracking data. In Section 3, we compare five DNN-based saliency algorithms and get the saliency model of autistic. Section 4 summarizes the whole paper and gets the conclusion.

## 2. EXPERIMENT AND ANALYSIS

### 2.1. Experiment

Nineteen high-functioning children with ASD were recruited as subjects. However, because of the difficult communication with the children with ASD, there were only thirteen subjects who could complete the calibration step and obtain relative effective eye-tracking data. The age of the remaining ASD participants ranged from 5 years old to 12 years old and the mean age of the subjects was 7.8 years old. All ASD participants had normal or corrected-to-normal visual acuity. All ASD participants met the diagnostic criteria of ASD in DSM-5 [14].

All of the 500 images were randomly collected from Judd *et al.* [11], which is a large database that contains images with a large variety of objects in or entirely natural scenes. With these different kinds of images, our database could help researchers understand the visual attention of children with ASD better from low-level feature contrasts to high-level semantic contrasts. The eye tracker we used is Tobii T120 with a 19 inch screen whose resolution is $1280\times1024$. The distance between the subjects and the eye tracker is 65 $cm$. Due to the lack of patience of ASD children, the experiment was split into ten recording sessions, with 50 randomly selected images in each session. At the start of each session, we re-calibrated the eye-tracker to ensure the reliability of the data. Each image was shown at full resolution for 3 seconds, with an one second of gray screen between two images. Subjects were told to look at the images freely, nevertheless, because of the lack of patience and difficulty concentrating, we had to remind them to look at the image. To conduct comparison experiment, we also collect eye-tracking data of healthy children as controls in the same way, and it is obviously easier than experimental group.

### 2.2. Analysis of visual attention map

In order to obtain a continuous fixation density map of an image from the eye tracking data of subjects, we overlay all fixation points of this image fixated by all viewers into one map, and then the map is smoothed with a Gaussian kernel (bandwidth = $1^\circ$) and normalized to a fixed dynamic range. Using these visual attention maps, we generate the fixation heat map of the images for autistic and healthy controls respectively. By comparing the similarity and the difference of heat maps between autistic and healthy controls, there are many conclusions which would help us improve the performance of the saliency model of autistic. We select some images from

**Table 1**. Performance of five state-of-the-art algorithms in our database (test set). The first line shows five algorithms respectively. In the second line, type 1 or 2 denote using original model of the algorithm or using fine-tuned model with autistic's fixation data, respectively. Line 3 to line 6 list the performance of corresponding evaluation results using the ground truth of autistic's fixation data while line 7 to line 10 list them using the ground truth of healthy controls' fixation data. The best performing model using each evaluation criterion is highlighted with bold.

| Algorithms | Salicon [7] | | SalGAN [8] | | mlnet [9] | | SAM-VGG [15] | | SAM-ResNet [15] | |
|---|---|---|---|---|---|---|---|---|---|---|
| Type | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 |
| ASD-AUC | 0.7476 | 0.7801 | 0.7855 | 0.8154 | 0.7680 | 0.7826 | 0.7769 | **0.8178** | 0.7807 | 0.8094 |
| ASD-sAUC | 0.5386 | 0.5610 | 0.5929 | **0.5951** | 0.5443 | 0.5806 | 0.5371 | 0.5636 | 0.5326 | 0.5682 |
| ASD-CC | 0.4437 | 0.5585 | 0.6064 | **0.7126** | 0.5254 | 0.6055 | 0.5594 | 0.7010 | 0.5529 | 0.6903 |
| ASD-NSS | 1.0826 | 1.2275 | 1.4114 | 1.5251 | 1.3062 | 1.3971 | 1.4040 | **1.5353** | 1.3866 | 1.4751 |
| Healthy-AUC | 0.8292 | 0.8335 | 0.8685 | 0.8638 | 0.8506 | 0.8606 | 0.8626 | 0.8698 | **0.8706** | 0.8474 |
| Healthy-sAUC | 0.5917 | 0.6120 | **0.6581** | 0.6455 | 0.6068 | 0.6491 | 0.5696 | 0.6231 | 0.5881 | 0.6166 |
| Healthy-CC | 0.5909 | 0.6075 | 0.7367 | 0.7264 | 0.6724 | 0.7044 | 0.6910 | **0.7568** | 0.7177 | 0.6897 |
| Healthy-NSS | 2.0124 | 1.7125 | 2.1148 | 1.9049 | 2.2672 | 2.0815 | 2.2169 | 2.0958 | **2.4382** | 1.8415 |

our database and show them with their corresponding fixation heat maps of autistic or healthy controls together in Fig. 1. There are six subfigures in Fig. 1. and the three columns in each subfigure are sample image, heat map of autistic and heat map of healthy controls from left to right, respectively. We will proceed the discussion of these figures subsequently.

The first three subfigures show a series of social activities and we analyse the difference between the autistic and healthy controls in three different aspects. Fig. 1. (a) shows the absence of joint attention of ASD. The healthy controls will scan the faces of different people and judge the relation from the sight (maybe instinctively), while autistic will concentrate more on the people or interesting objects in the center zone of the this kind of images without the consideration of joint attention. The image center bias described by Wang *et al.* [5] would be more obvious when there is hardly any joint attention information, just like the situation in Fig. 1. (c). However, it is interesting that the absence of joint attention of ASD will disappear when the target is object or animal and they even pay more attention to these areas of images. The normal controls will also fixated more on human faces in this kind of situation. Thus we get the difference of social bias of ASD.

Fig. 1. (d) shows another different bias which is named as hand bias. In the situation where there is an interactive activity between objects and the hand of the main character, autistic will pay more attention to the hand and the objects in the hand. This is an interesting and fairly useful phenomenon which could guide researchers to design textbook related to ASD. In Fig. 1. (e), we observe a new phenomenon which against the center bias described in [5]. For natural scenes, ASD will fixated more on pixel level features and the distribution of autistic's fixation point is very scattered. On the contrary, the visual attention maps of these images collected from healthy controls have an obvious center bias. Therefore, we suppose that the center bias is related to whether subjects interest in the image. If they do not have interest in the image, they will have a more obvious center bias. With the consideration of human faces, we show three images with human faces in Fig. 1. (f). It is obvious that the two visual attention maps fixated by autistic and healthy controls are similar though there are tiny differences. Therefore, when faces are the main contents of the image, the difference between autistic's visual attention map and healthy people's visual attention map is slight.

There are many other features of autistic's visual attention map, such as more concentrate on animals, mouth bias, even sometimes they will look arbitrarily. Our database will be released to facilitate further research and the content of database will be more and more abundant.

## 3. LEARNING TO PREDICT WHERE THE CHILDREN WITH ASD LOOK

The goal of establishing the database is to predict where the children with ASD look. So based on the analysis about the accordance or difference between autistic's visual attention map and healthy controls' visual attention map, we transform the existing saliency prediction models of healthy people to the children with ASD. In this paper, we compare and fine-tune five state-of-the-art algorithms for image saliency prediction based on our database.

There are 500 images in total in the database and we randomly select 400 images as training group and the rest of the images are assign to test group. because the amount of the proposed database until now is inadequate to train an end-to-end deep neural network due to over-fitting, we decide to fine-

**Table 2**. Comparison of the performance with the model fine-tuned by autistic's fixation data and healthy controls' fixation data respectively. Type 2 represents results fine-tuned by autistic's fixation data. Type 3 represents results fine-tuned by healthy controls' fixation data

| SALICON | SalGAN | | SAM-VGG | |
|---|---|---|---|---|
| Type | 2 | 3 | 2 | 3 |
| ASD-AUC | 0.8154 | 0.8043 | **0.8178** | 0.8063 |
| ASD-sAUC | **0.5951** | 0.5909 | 0.5636 | 0.5503 |
| ASD-CC | **0.7126** | 0.6649 | 0.7010 | 0.6549 |
| ASD-NSS | 1.5251 | 1.4586 | **1.5353** | 1.5123 |
| Healthy-AUC | 0.8638 | **0.8843** | 0.8721 | 0.8818 |
| Healthy-sAUC | 0.6455 | **0.6794** | 0.6231 | 0.6159 |
| Healthy-CC | 0.7264 | **0.8098** | 0.7568 | 0.7927 |
| Healthy-NSS | 1.9049 | 2.1742 | 2.0958 | **2.3677** |

tune the provided networks and transfer these models from normal people to people with ASD. However, as the expansion of the database, we could train the model rather than fine-tuning it in the future. The architecture of the network that Jiang *et al.* used the in [6] follows the design of the SALICON network [7], while the SALICON network consists of two parallel VGG-16 networks [16] which process the input images at two different spatial resolutions. In this paper, we also use the SALICON as one of five algorithms for its good performance in saliency prediction of healthy people. The parameters of the SALICON network are same with that in [7]. However, to get the best method of predicting where the children with ASD look, we fine-tune another four state-of-the-art algorithms including SalGAN [8], mlnet [9], SAM-VGG and SAM-ResNet [10] and then compare them. The learning rates of SalGAN, mlnet, SAM-VGG and SAM-ResNet are changed to $10^{-3}$, $10^{-2}$, $10^{-6}$ and $10^{-6}$ respectively to get better performance when maintaining other parameters consistent with the original settings. The networks are then fine-tuned on the current database.

Four evaluation criteria, including AUC-Judd, sAUC, CC, NSS [10, 17, 18, 19, 20] are used to evaluate the performance of the saliency models of autistic for the first time. Detailed results are listed in Table 1. In the first line of the table, we list the five algorithms respectively, including Salicon, SalGAN, mlnet, SAM-VGG and SAM-ResNet. Type 1 or type 2 in the second row of the table means using original model of the algorithm or using fine-tuned model with autistic's fixation data, respectively. Line 3 to line 6 list the performance of five algorithms with two different types of models using the ground truth of autistic's fixation data. Line 7 to line 10 shows the performance that using fixation data of healthy controls as ground truth. The best performing algorithm under each evaluation criteria is highlighted with bold.

As shown in the third line to the sixth line, all of four algorithms including SalGAN, mlnet, SAM-VGG and SAM-ResNet have better performance than SALICON when using autistic's fixation data as ground truth no matter before or after fine-tuning. And it is obvious that after fine-tuning, the performance of all algorithms under four evaluation criteria has a great degree of promotion. The two best performing algorithms are SalGAN and SAM-VGG. The performance is not impressive but also not too bad. From line 7 to line 10, when considering four evaluation criteria together, we could find that the performance are invariant or even decline when the fixation data of healthy people serves as ground-truth. However, two criteria, sAUC and CC almost always increase, though the magnitude is smaller than above. So we also conduct an comparison experiment and fine-tune the network with fixation data of healthy controls. We select two best performing algorithms on our database and show the comparison results in Table 2.

From Table 2, we can see that, from line 3 to line 6, the performance of the network fine-tuned by autistic's fixation data is better than that of the network fine-tuned by healthy controls' fixation data. The situation is on the contrary for line 7 to line 10. Since the fine-tuning parameters are the same in both two cases, we can say that we get a relatively good saliency model for autistic, though this model is still closer to healthy controls.

The fine-tuning step only changes the layers' weights of the network, and in this paper, we believe that the layers related to pixel levels will obtain more weights after fine-tuning. In the future, we will consider changing the structure of the network and combining features we provide in Section 2.2 to get better saliency model of autistic. The database will be released and welcome other researchers to try their ideas based on our database.

## 4. CONCLUSION

In this paper, we first conduct an experiment to get the fixation data of autistic and then establish a database related to saliency prediction for children with autism, named SPCA. Based on the database, many features which could be used to train autistic's saliency model are shown, including absence of joint attention, more concentrate on objects or animals, center bias, hand bias, against center bias under natural scenes and so on. We compare five state-of-the-art saliency prediction models and then fine-tune these models using the proposed database. Experimental results indicate that we get a relatively good saliency model suitable for autistic. In the future, we will enlarge our database and improve the structure of the network based on the features described above to train a better saliency model for children with autism.

# Acknowledgment

This work was supported by the National Science Foundation of China (61521062, 61527804) and Science and Technology Commission of Shanghai Municipality (15DZ0500200).

## 5. REFERENCES


[1] Geraldine Dawson, Sara Jane Webb, and James McPartland, "Understanding the nature of face processing impairment in autism: insights from behavioral and electrophysiological studies," *Developmental neuropsychology*, vol. 27, no. 3, pp. 403–424, 2005.

[2] Julie Osterling and Geraldine Dawson, "Early recognition of children with autism: A study of first birthday home videotapes," *Journal of autism and developmental disorders*, vol. 24, no. 3, pp. 247–257, 1994.

[3] Po-He Tseng, Ian GM Cameron, Giovanna Pari, James N Reynolds, Douglas P Munoz, and Laurent Itti, "High-throughput classification of clinical populations from natural viewing eye movements," *Journal of neurology*, vol. 260, no. 1, pp. 275–284, 2013.

[4] Wenbo Liu, Ming Li, and Li Yi, "Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework," *Autism Research*, vol. 9, no. 8, pp. 888–898, 2016.

[5] Shuo Wang, Ming Jiang, Xavier Morin Duchesne, Elizabeth A Laugeson, Daniel P Kennedy, Ralph Adolphs, and Qi Zhao, "Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking," *Neuron*, vol. 88, no. 3, pp. 604–616, 2015.

[6] Ming Jiang and Qi Zhao, "Learning visual attention to identify people with autism spectrum disorder," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3267–3276.

[7] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao, "Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 262–270.

[8] Junting Pan, Cristian Canton, Kevin McGuinness, Noel E. O'Connor, Jordi Torres, Elisa Sayrol, and Xavier and Giro-i Nieto, "Salgan: Visual saliency prediction with generative adversarial networks," in *arXiv*, January 2017.

[9] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara, "A Deep Multi-Level Network for Saliency Prediction," in *International Conference on Pattern Recognition (ICPR)*, 2016.

[10] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand, "What do different evaluation metrics tell us about saliency models?," *arXiv preprint arXiv:1604.03605*, 2016.

[11] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba, "Learning to predict where humans look," in *IEEE international conference on Computer Vision*. IEEE, 2009, pp. 2106–2113.

[12] Jonathan Harel, Christof Koch, and Pietro Perona, "Graph-based visual saliency," in *Advances in neural information processing systems*, 2007, pp. 545–552.

[13] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal, "Context-aware saliency detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 10, pp. 1915–1926, 2012.

[14] American Psychiatric Association et al., *Diagnostic and statistical manual of mental disorders (DSM-5®)*, American Psychiatric Pub, 2013.

[15] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara, "Predicting human eye fixations via an lstm-based saliency attentive model," *arXiv preprint arXiv:1611.09571*, 2016.

[16] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[17] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba, "Mit saliency benchmark," .

[18] Xiongkuo Min, Guangtao Zhai, Ke Gu, and Xiaokang Yang, "Fixation prediction through multimodal analysis," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 13, no. 1, pp. 6, 2017.

[19] Xiongkuo Min, Guangtao Zhai, Ke Gu, Jing Liu, Shiqi Wang, Xinfeng Zhang, and Xiaokang Yang, "Visual attention analysis and prediction on human faces," *Information Sciences*, vol. 420, pp. 417–430, 2017.

[20] Ke Gu, Guangtao Zhai, Weisi Lin, Xiaokang Yang, and Wenjun Zhang, "Visual saliency detection with free energy theory," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1552–1555, 2015.